MACHINE LEARING

FIRST ASSIGNMENT

You have a free choice of programming language.

Work in groups. The optimal number of students in a group is 3. You can earn bonus points for sharp comments and creative thinking.

Deadline: Tuesday, 11 January 2022

Problem 1

Consider the generalized linear regression problem defined by the following model:

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_5 x^5 + \eta \qquad (1)$$

where η corresponds to white Gaussian noise and the components of the weight vector assume the following values:

$$\theta_0 = 0.2\,,\ \theta_1 = -1\,,\ \theta_2 = 0.9\,,\ \theta_3 = 0.7\,,\ \theta_5 = -0.2\,. \qquad (2)$$

In every case below, we consider N equidistant points $x_1, x_2, ..., x_n$ in the interval [0,2] and use them to create samples for our training set:

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 x_n^2 + \theta_3 x_n^3 + \theta_5 x_n^5 + \eta_n, \quad n = 1, 2, ..., N \qquad (3)$$

where η_n are i.i.d. noise samples originating from a Gaussian distribution with mean 0 and variance σ_η^2.

1) Using $N = 20$, σ_η^2 =0.1 and the structure of the correct model (5th degree polynomial with the coefficient of the 4th power equal to zero), apply the Least Squares method to estimate the parameter vector. Calculate the Mean Square Error of *y* over the training set and over a test set comprising of 1000 points randomly selected in the interval [0,2].
2) For $N = 20$ and σ_η^2 =0.1 apply regression using the Least Squares method and a 2nd degree polynomial. Perform 100 experiments using different noise samples for each experiment. For each point of the training set, calculate the mean and variance of *y* over the 100 experiments and plot these quantities on the (*x,y*) plane along with the curve obtained by the true model.
   Repeat using a 10th degree polynomial. Compare your results obtained for the 2 different cases (2nd versus 10th degree polynomial) making special reference to the bias-variance dilemma.
3) Repeat experiment (1) above, implementing the Ridge Regression method with various values of λ (instead of the Least Squares Method). Report whether you have observed an improvement of the Mean Square Error for some of the values of λ.
4) We encode our prior knowledge for the unknown parameter vector via a Gaussian distribution G(**θ**) with mean $\boldsymbol{\theta}_0$ equal to the true parameter vector in equation (1) and covariance matrix $\Sigma_\theta = \sigma_\theta^2 I,\ \sigma_\theta^2 = 0.1$. Use the structure of the true model and

perform full Bayesian Inference in order to evaluate *y* for 20 randomly selected test set points belonging to the interval [0,2] and for two different values of σ_η^2 (0.05 and 0.15). Plot your estimates and their errors on the (*x,y*) plane.

5) Repeat experiment (4) using the following mean vector for G(**θ**): $\boldsymbol{\theta}_0 = [-10.54, 0.465, 0.0087, -0.093, -0.004]^T$

   With σ_η^2 =0.05, perform the experiment four times, using two different values for σ_θ^2 (0.1 and 2) and two different values for N (20 and 500). Comment on your results.

6) Try to recover the true variance of the noise using the Expectation-Maximization method. Construct a training set with $N = 500$ and σ_η^2 =0.05. Initialize the algorithm with $\alpha = \sigma_\theta^{-2} = 1, \beta = \sigma_\eta^{-2} = 1$. After convergence, estimate the *y*'s and their errors over a test set of 20 points randomly selected in the interval [0,2]. Plot these quantities on the (*x,y*) plane, along with the true model curve.

Problem 2

1) Program and implement a k nearest neighbours classifier (k-NN). Use this classifier to solve the following problems:
   i. IRIS PLANT DATABASE (Classification of three different kinds of iris plants).
   ii. PIMA INDIANS DIABETES DATABASE (Classification of pregnant Indians of the Pima tribe according to whether they suffer from diabetes or not).

   The relevant data can be found in the file UCIdata-exercise1.rar.
   Report on the percentage of correct classification as a function of the number of nearest neighbours. Use cross-validation to obtain the results.
2) For the second problem, obtain estimates of the probability density functions for each class, under the following assumptions:
   a) Pdfs are gaussian. The covariance matrices are diagonal, with all diagonal elements equal. Mean and variance of the pdfs are estimated using Maximum Likelihood from the available data.
   b) Pdfs are gaussian, with non-diagonal covariance matrices. Means and covariance matrices of the pdfs are estimated using Maximum Likelihood from the available data.
   c) Components of the feature vectors are mutually statistically independent (the usual naïve Bayes approach). Marginal Pdfs are gaussian, with parameters (mean, variance) estimated using Maximum Likelihood from the available data.
   d) Components of the feature vectors are mutually statistically independent (the usual naïve Bayes approach). Marginal pdfs are computed using 1-d Parzen windows with gaussian kernels. Take the width *h* of each window equal to the square root of the number of patterns in the available data.

   For all assumptions compute the following measures of the goodness of your fit for each class: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) (https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118856406.app5). For this

question, use the whole data set available to compute the above measures (do not use cross validation).

3) For each of the above assumptions about the pdfs, implement a Bayes classifier and compute its classification accuracy using cross validation (it goes without saying that for each cross validation iteration, probability density functions will have to be calculated using only training set data for this question). For assumptions c) and d), this will obviously be a naïve Bayes classifier. Taking into account your previous findings, investigate whether more accurate estimates for the pdfs (as judged by the model selection criteria in question 2) tend to improve classification accuracy as well. Compare the performance of the Bayes classifiers to the performance of the k-NN classifier.
4) Implement the perceptron algorithm and use it to perform classification on the IRIS PLANT DATABASE data as follows: Examine whether the data of each class are linearly separable from the data of the combined remaining classes (e.g. if the Iris Setosa data are linearly separable from the combined Iris Versicolor and Iris Virginica data).